REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
01/07/00

3. REPORT TYPE AND DATES COVERED
Scientific/Tech 10/01/99 - 12/31/99

4. TITLE AND SUBTITLE
Browsing, Discovery, and Search in Large Distributed Databases of Complex and Scanned Documents

5. FUNDING NUMBERS
F19628-95-C-0235
ARPA Order No. D570

6. AUTHOR(S)
W. Bruce Croft

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Massachusetts, Amherst
Box 36010, OGCA, Munson Hall
Amherst, MA 01003-6010

8. PERFORMING ORGANIZATION REPORT NUMBER
TR5281811299

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Mr. Charles Shank
ESC/PKRB
104 Barksdale St., Bldg 1520
Hanscom AFB, MA 01731-1806

Ms. Monique Dillon
Office of Naval Research
Boston Regional Office
495 Summer St., Room 103
Boston, MA 02210-2109

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This project aims to integrate powerful, new techniques for interactive browsing, discovery, and retrieval in very large, distributed databases of complex and scanned documents. Emphasis is placed on going beyond full-text retrieval techniques developed in the DARPA TIPSTER program to support different types of access and non-textual content. These techniques should be particularly relevant to the patent domain where it is important to find relationships between documents and where the patent or trademark may be based on a visual design. The specific tasks identified involve studying representation techniques for long documents with complex structure, browsing and discovery techniques for large text databases, image retrieval and scanned document retrieval techniques, and architectures for large, distributed databases.

20000112 009

14. SUBJECT TERMS
Browsing
Image Retrieval
Text Retrieval
Query Processing
Scanned Document Retrieval
Probabilistic Retrieval Model
Indexing
Bayesian Network
Large Distributed Databases

15. NUMBER OF PAGES
11

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Scientific and Technical Report

Sponsored by
Advanced Research Projects Agency/ITO
and United States Patent and Trademark Office

Browsing, Discovery and Search in Large Distributed Databases
of Complex and Scanned Documents

ARPA Order No. D570

Issued by EXC/AXS under Contract #F19628-95-C-0235

Date Submitted: January 7, 2000

Period of Report: October 1, 1999 to December 31, 1999

Submitted by: Professor W. Bruce Croft, Principal Investigator
Computer Science Department
University of Massachusetts, Amherst

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government.

Distribution Statement A: Approved for public release; distribution is unlimited.

Table of Contents

Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents

Technical and Scientific Report

Task 1: Representation Techniques for Complex Documents

Task Objectives

In this task, the goal is to extend the word-based representations that are common in retrieval systems in order to support summarization, browsing, and more effective retrieval. Specifically, we have been studying phrase-based representations and relationships between phrases in individual and groups of documents as the basis for our approach. Document structure will be used as part of the information that is used to "tag" the phrasal representation.

Technical Problems

The technical problems have to do with defining a "phrase", developing techniques for rapidly extracting them from text, comparing phrase contexts to identify significant relationships, producing summaries from these representations, extending the underlying retrieval model to be able to make effective use of phrasal representations, and using complex document structure in indexing and retrieval.

General Methodology

The general methodology for this task is to demonstrate effectiveness through user-based and collection-based experiments. As well as the PTO text databases, we will make extensive use of the TIPSTER document collection, which consists of a large number of text documents from a variety of sources, queries, and user relevance judgments for each query.

Technical Results

We are concluding our research on how best to use phrases in a query containing a sequence of words that forms a phrase. Our research has shown that when the single terms are high frequency terms and they co-occur in documents frequently, but the phrase (two terms in proximity) occurs infrequently, it is better to use the phrase. We completed exploratory data analyses on the statistical distributions of phrases and the terms that make up the phrases, in whole collections, and in relevant documents, found that phrases have similar *idf* (inverse document frequency) distributions to single terms, so that the best way to weight phrases in queries is by their *idf* weights. The following papers report this work and other exploratory data analyses:

- Pickens, J. and Croft, W.B. "An Exploratory Analysis of Phrases in Text Retrieval," submitted to RIAO 2000 Conference, Paris, France, April, 2000.

- Greiff, Warren R. "The Use of Exploratory Data Analysis in Information Retrieval Research," to appear *in Advances in Information Retrieval: Recent Research from the CIIR*, W. Bruce Croft, ed., Kluwer Academic Publishers, 2000.

We have continued to develop the new retrieval technique based on language models. We continue to carry out experiments to see the effects of learned language models on retrieval.

- Ponte, Jay. "Language models for relevance feedback" to appear *in Advances in Information Retrieval: Recent Research from the CIIR*, W. Bruce Croft, ed., Kluwer Academic Publishers, 2000.

- Croft, W. Bruce. "Combining Approaches to Information Retrieval" to appear *in Advances in Information Retrieval: Recent Research from the CIIR*, W. Bruce Croft, ed., Kluwer Academic Publishers, 2000.

- Xu, Jinxi, and Croft, W. Bruce. "Topic-Based Language Models for Distributed Retrieval" to appear *in Advances in Information Retrieval: Recent Research from the CIIR*, W. Bruce Croft, ed., Kluwer Academic Publishers, 2000.

Important Findings and Conclusions

Language models are becoming an important in modern search engines.

Significant Hardware Development

None.

Special Comments

None.

Implication for Further Research

Language models may provide retrieval improvements for PTO data.

Selective incorporation of phrase operators with appropriate weights may improve retrieval on PTO queries.

Task 2: Browsing and Classification Techniques for Document Collections

Task Objectives

The goals of this task are to develop techniques for summarizing and classifying collections of documents. These techniques will be designed to support interactive browsing and text classification in environments like the PTO.

Technical Problems

The technical problems involve producing an effective summary of a group of documents, such as a retrieved set or an entire database. Both document and phrase clusters could be used as part of this process. The classification task emphasizes the ability to accurately assign predefined categories (as in the PTO classification) to new documents (patents). An additional problem is to determine when existing classifications do not match well to new documents, such as when a PTO category covers too many patents and needs to be refined.

General Methodology

Evaluation of these techniques will be done using both the TREC corpus and PTO data. For the classification task in particular, we are designing evaluation criteria with substantial input from PTO staff.

Technical Results

The results from earlier distributed retrieval research showed that collection selection is an effective method for dealing with collections that must be subdivided into smaller collections. We are, therefore, working on a multi-level classification scheme which involves dividing a database into smaller databases by class and then using collection selection to find the appropriate class. We are comparing standard classification algorithms with collection selection algorithms for the first-level classification scheme.

We are continuing work with Dataware Technologies to evaluate different approaches to classification.

In the summarization/visualization area, we have continued to develop techniques for combining ranked lists with clustering. A ranked list is a well-known technique for presenting information so that relevant documents may be found quickly. Clustering is also a well-known technique for grouping similar documents. We have improved our method of combining the two approaches and developed a better evaluation technique, providing more effective and robust results. This work is described in the following paper:

- Leuski, A. and Allan, J. "Improving Interactive Retrieval by Combining Ranked List and Clustering," submitted to RIAO 2000 Conference, Paris, France, April, 2000.

Also in the summarization/visualization area we have done further work on deriving Yahoo-like subsumption hierarchies that extract significant relations among words related to a query. Display of these hierarchies provides a query-sensitive "summary" of the content of a collection.

- Lawrie, D. and Croft, W.B. "Discovering and Comparing Hierarchies," submitted to RIAO 2000 Conference, Paris, France, April, 2000.

- Sanderson, Mark, and Lawrie, Dawn. Building, Testing, and Applying Concept Hierarchies. to appear *in Advances in Information Retrieval: Recent Research from the CIIR*, W. Bruce Croft, ed., Kluwer Academic Publishers, 2000.

Important Findings and Conclusions

None.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

These visualization and summarization techniques could enhance the user interface of a search and classification system such as that of the PTO demo. We are using the multi-level classification scheme in the demo system, and will update it to include the best performing classification approach for the first stage. We are still evaluating different approaches to classification with Dataware.

Task 3: Image Indexing and Retrieval

Task Objectives

The goal of this task is to develop similarity-based techniques for retrieving images such as trademarks, logos, and designs.

Technical Problems

The central issue is how images can be indexed to support efficient, content-based retrieval. The primary type of query in these environments is "find me things that look like this". We are developing "appearance-based" retrieval of images as well as more straightforward features such as color and texture. Filter based and frequency domain based techniques offer some potential in this area, but significant work needs to be done on making this approach efficient enough to deal with hundreds of thousands of images.

General Methodology

The evaluation of these techniques will be done in a similar way to text by developing test collections of images. Specifically, we are working to obtain large collections of trademark and design images, both from the PTO and from general sources such as the web.

Technical Results

Our work is now focused on evaluating the demonstration system. We have compared the results of our image retrieval system and a technique based on invariant moments (a technique used by some researchers to retrieve trademarks). The techniques were first compared on a set of synthetic geometric images created for the purpose of testing. On this set, the curvature-phase technique developed by us far outperformed the moment based technique (74.8% versus 25.6%). We then tested it on a collection of geometric trademarks with relevance judgments from the British Patent Office (The relevance judgments were performed by a trademark examiner). On this set, also the curvature-phase technique performed better than the moment based technique. This work will appear in the following paper:

- Ravela, S. and Luo, C. "Appearance-Based Global Similarity Retrieval of Images", To appear in "Advances in Information Retrieval", Ed. W. B. Croft, Kluwer Academic Publishers, 2000.

On closer examination, we discovered that the relevance judgments for the British trademarks were not completely reliable. We are, therefore, working on creating our own relevance judgments based on visual similarity. For this purpose, we have created a user interface. We also continue to improve the effectiveness of our trademark retrieval system.

We have collected 4000 additional flower images from the web. We are continuing to index these images and adding them to the flower patent database to judge retrieval effectiveness over a larger set. This work on flower patent images has now appeared in a journal:

- Das, M, Manmatha, R. & Riseman, E. M. “Indexing Flower Patent Images Using Domain Knowledge”. In IEEE Intelligent Systems, Sept/Oct 1999, pages 24-33.

We continue our work on modifying our technique for segmenting flowers from images may also be used for segmenting other objects like birds. Much more work needs to be done in this area, but the initial work shows this to be a promising approach for segmenting objects from (certain kinds of) images so that they can be further indexed for retrieval.

Our work on detecting text in images has also appeared in a journal:

- Wu, V, Manmatha, R. & Riseman E. M. "TextFinder: An Automatic System to Detect and Recognize Text in Images”. In IEEE Transactions of Pattern Analysis and Machine Intelligence, Nov 1999, pages 1224-1228.

Important Findings and Conclusions

None.

Significant Hardware Development

None

Special Comments

None.

Implications for Further Research

We continue to focus on evaluating the accuracy of our techniques using trademark testbeds from Britain and, hopefully, from the U.S. PTO. We will also continue to improve the demonstration system.

Task 4: Distributed Retrieval Architecture

Task Objectives

The goals of this task are to scale up our current methods of automatically selecting collections and merging results, and to investigate architectures that can support efficient retrieval, browsing and relevance feedback in distributed environments with terabytes of information.

Technical Problems

The current INQUERY text retrieval system uses a client-server architecture to support simultaneous retrieval from multiple collections distributed across one or more processors. A number of efficiency bottlenecks develop, however, when the size of the databases is very large. Deciding which subcollections to search can address part of the problem, but there are other problems associated with the fundamental efficiency of the processes involved and the use of distributed resources. Image indexing and retrieval tends to exacerbate these efficiency problems since the databases and indexes are considerably larger.

General Methodology

The architectures and algorithms produced in this task will be evaluated using a combination of standard performance (efficiency) measures and effectiveness measures. The efficiency tests will be done using TREC data and large PTO databases, including images, and the collection selection algorithms will be evaluated using the text subcollections of the patents.

Technical Results

We have continued to extend our work on query-based sampling to build models of multiple collections in several directions. Earlier work showed that there was no noticeable loss of in collection-selection effectiveness when using models obtained by sampling vs complete models of each collection. An overview of this and newer work can be found in:

- Callan, Jamie. "Distributed Information Retrieval" to appear *in Advances in Information Retrieval: Recent Research from the CIIR,* W. Bruce Croft, ed., Kluwer Academic Publishers, 2000.

We have extended the work on sampling by studying the generality and effectiveness of query-based-sampling with different database selection algorithms, and with queries of varying length and accuracy, reported in the following paper:

- Callan, J., Powell, A., French, J. and Connell, M ., " The Effects of Query-Based Sampling on Automatic Database Selection Algorithms," submitted to SIGMOD, Dallas, TX, May 15-19, 2000.

Results from other research suggest that it if large collections need to be divided into smaller collections, it is better to organize by subject. We are verifying this with PTO data and comparing four different algorithms for merging databases to determine whether different merging algorithms are optimal for different database organizations.

We have continued to explore architectural issues concerning subdivided collections in:

- Lu, Zhihong, and McKinley, Katheryn S. "The Effect of Collection Organization and Query Locality on Information Retrieval System Performance" to appear *in Advances in Information Retrieval: Recent Research from the CIIR*, W. Bruce Croft, ed., Kluwer Academic Publishers, 2000.

Important Findings and Conclusions

Organization by topic is likely to be better than a chronological organization. Distributed search can be more effective than centralized search if it is based on language models. Database sampling is the first practical method of discovering (or verifying) the contents of databases controlled by third parties.

Significant Hardware Development

None.

Special Comments

None

Implications for Further Research

Organizing documents by subject is likely to be more effective than organizing them by the date of the document. We will continue to evaluate performance of distributed architectures for scalable IR using the new demonstration system. Database sampling is a good practical way of discovering (or verifying) the contents of databases controlled by third parties.